GUSTO Data Vault: Laying the foundations for an open science system with OMOP Data Catalogue

Cindy Ho^{1,2}, Li Ting Ang^{1,2}, Maisie Ng^{1,2}, Hang Png^{1,2}, Shuen Lin Tan^{1,2}, Estella Ye^{1,2}, Sunil Kumar Raja¹, Mengling Feng^{3,4}, Johan G Eriksson^{1,5,6,7}, Mukkesh Kumar^{1,2,3}

Institution(s) of origin:

¹ Singapore Institute for Clinical Sciences, Agency for Science Technology and Research, Singapore

² Bioinformatics Institute, Agency for Science Technology and Research, Singapore

³ Saw Swee Hock School of Public Health, National University of Singapore, National University Health System, Singapore

⁴ Institute of Data Science, National University of Singapore, Singapore

⁵ Department of Obstetrics and Gynaecology and Human Potential Translational Research

Programme, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

⁶ Department of General Practice and Primary Health Care, University of Helsinki, Helsinki, Finland

⁷ Folkhälsan Research Center, Helsinki, Finland

Background

The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) created by Observational Health Data Sciences and Informatics (OHDSI), is an open community data standard, designed to standardise the structure and content of observational data¹. The harmonisation of data using OMOP CDM enables systematic and collaborative OMOP-based research. The Agency for Science, Technology and Research (A*STAR) has been managing the datasets for national level research programmes such as the Growing Up in Singapore Towards healthy Outcomes (GUSTO)² birth cohort study. A*STAR has pioneered health research with an open interactive GUSTO Data Vault platform (Figure 1) for hypothesis construction and data-driven discoveries^{3,4}. The OMOP Data Catalogue in GUSTO Data Vault showcases the data assets which have been converted into Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) format. The current CDM version is CDM v5.4⁵, as depicted below for GUSTO tables and fields (Figure 2).



Figure 1. GUSTO Data Vault Homepage



Figure 2. OMOP CDM Model for GUSTO

Methods

Reinventing the working culture around OMOP data literacy, research organisations and pharmaceutical companies can create a data governance structure with OMOP Data Catalogue. The OMOP Data Catalogue was developed using open-source technologies. The containerised web application with Docker was built using PostgreSQL database engine and Django web application framework. The GUSTO Data Vault application is deployed on AWS cloud environment. The front-end (client side) scripting is using HTML, CSS, jQuery, Ajax. The back-end (server side) scripting is using Python. The GUSTO OMOP fields showcased in OMOP Data Catalogue were mapped using open source OMOP tools provided by OHDSI and R programming community.

Results

The FAIR (Findable, Accessible, Interoperable, Reusable) guiding principles for data management and stewardship is a critical need for academia, industry, funding agencies and scholarly publishers⁶. The OMOP Data Catalogue makes GUSTO cohort-specific CDM fields to be discovered by the global research community. The OMOP CDM fields can be discovered across the Person, Condition, Observation and Measurement tables (Figure 3). Researchers can perform search across variables and concepts. The metadata is described with a plurality of relevant attributes such as CDM Field, Concept ID, Concept Type, Subject Type, Visit Timepoint, Description and Domain. The data profiling of the OMOP Concept IDs enables GUSTO data to be identified, described, discovered, and reused by researchers (FAIR data principles). The visual interface of OMOP Data Catalogue guides researchers through the GUSTO tables and fields. The OMOPed data from incremental OMOP conversions (e.g. Table-by-table conversions) can be seamlessly integrated in OMOP Data Catalogue by GUSTO data curators.

Home Data Catelogues QUEID OMOP Common Data Model	Emperimana Device	Data Registrand Data 1	Disaduations Publications Request for Data Search Data VedR													
	GUSTO O	MOP Common	Data Model	nershia-IOMC	P) Common D	Home Outs Catalogues GUSTO CMOP Commo	on Data Model	lispecinens	Onies Data	Playground Data Vouslisations	Publications Request for Da	as Search Data Vavit				
Cirical Data	Nodel (CDM) form	at OMOP COM is an open of	mmunity data standard, designed to standard se the structure and contest of observ	otonal data."	The	n Rouma	1 =									
Person	GUSTO tables and I	ahe using CMOP CDM enable Selds.	is systematic and collaborative GMOP-based research. The current CDM version is C	ow which an de	quicked below)			Observation								
Condition						Home		Subject Ty:				0				
		Design (B Method) (C) (See	Displaying	Clinical Data		Visit Timey	Description	along Briding		× WEZ × WEB × WEBB ×						
Other water		Chervator pe	visation visation					Keywords	Mother's pla	er's place of birth (self-reported) at recruitment.						
Measureneed		· Visit_occurren	Corcege			Petition			Concept ID	Description	Solgert Count (N)					
		1 . v	Vocabulary			Candition		Concept I	4152229	Born in Singapore	914 (64,69%)	Timpoint	Rolated Concepts			
	8	Canditoe_act	surence Domain			Observation		4264419	4152222	Born in Malaysia	104 (11.01%)	PATT NA WE				
	1	Measurem	Cancept agrowth			Measurement		4000105	4212913	Born in China	171 (12,1%)	PATT				
	0.1	Observat	Kin Standardond Cencept_relationality					44729990	4153168	Borninindia	115-07.4336	CH.				
	a diama d	Note	Relationship COM_Bource					20000000	4212175	Born in Argentina	1 (0.07%)	C01.	0			
	6	Fe Spectree	Source_to_concest_map Netsiasa					4331696	415212	terr in lines from	10275	PWTT	0			
										Born in Indonesia	29-02.09%0					
								4264549	4151606	Born in Ireland	1.0.07%)	Pani	0			
								4155450	4154061	Born in Kowait	1 (0.07%)	PMTI	0			
Clinical Bata •	CMD THEM	Description		Domain	Related Concepts											
Condition	penon_id	Subject ID of the Person				Home Data Catalogues GUSTO CM/CP Conve	on Data Model	Dospeckrens	Onice Data	Payground Data Visualizations	Publications Request for Dr	eta Search Data Vault				
Observation	gener/conopt_d	Bological ses at birth of	the Person.	Gender	0		÷ =									
Measurment	year, or, pers	March of birth darked from	n date of birth of the fersion			C SGUSTO										
	Instance/Look Subscription-Subscription and in Look of the Meson dig_LCloth Dig of Left dama Chan date of Look of the Meson sub_participation of the Meson (Sel Argonit of Meson/Selection of Selection					Harry	Date:				Measurement					
							(Selection)			Seinet All 😨 🛔 Victure 😨 🍈 Child						
								visit Timepoi	en l	PMITE PMUES DELS MIZX MIEX VEEX VEEX VEEX VEEX VEEX						
						Person		nywith								
						Condition		Cancept ID	Subject Type	Concept Type	Description	Wait Timpoints				
Home Data Catalogues 02/520 OMOP Converse Data Model	Respectivers Orrica	Data Planninand Data Y	Sculingform Dublications Research for Data Search Data Useds			Observation		3089413	Noter	Insulin measurement	Fasting Insulin reading	VM 103				
ă						Measurement		3017338	a Mather	Ourose measurement, quantificitive	2-hour plasma glucesa level	7418				
								3017538	& Noter	Gucce measurement, quantificative	2 hour OGTT reading	V64 (00) (V6				
	Condition							3018251	a Mother	Glucos measurement, quatilitative	Fasting plasma plucose	PM16				
Home	Subject Type	🛃 Select All 🛃 🛃	Mother					3018251	ANoter	Guose reasurement, quantitative	Facting OSTT reading	V64 (HS V68				
Cinical Data	Visit Timepoints	PHDS × DEL ×						3026145	a Mother	Insulin measurement	2 hour Insulin reading	YM INS				
Person	Reywords							4016239	Allother	Red blood cell indices determination	Mother's Mean Cell Volume	VHA INS				
Constant of Consta								4030871	a Nother	Blood cell count	Mother's Red Dicod Cell	VNA INS				
Constant of the second s	Concept Sub 10 Ty	pet Name	Description	timepoint	Concepts											
Observation	4024559	Cestational diabetes	Gestational Diabetes Mellinas (SDM) classification by WHO 1999 guidelines dusting PG +=	Paids	0											
Measurement		mellius	List of a more than 2 + 12.0													
	4290201	Hinding of beth outcome	e awar oussame or o.0570 pregnancy	DEL												
	4125390 🛔 Mr	ther delivery	Mother's Delivery mode	DEL	0											

Figure 3. OMOP Tables for GUSTO

Conclusions

The database schemas can vary for population health studies and involves considerable data engineering efforts for data standardisation. Singapore have initiated national level OMOP mapping to curate clinical datasets in OMOP standard to improve data interoperability. The GUSTO OMOP Data Catalogue developed for a Singapore cohort study lays the foundations for developing cross-study OMOP Data Catalogues expanded across APAC and global OHDSI data partners, enabling database level characterizations for knowledge discovery and management in an open science system.

References/Citations

- 1. Observational Health Data Sciences and Informatics. The Book of OHDSI [Internet]. [place unknown: Observational Health Data Sciences and Informatics]; 2021 [cited 2023 Apr 10] Available from: <u>https://ohdsi.github.io/TheBookOfOhdsi/</u>.
- Soh S-E, Tint MT, Gluckman PD, Godfrey KM, Rifkin-Graboi A, Chan YH et al. Cohort profile: Growing Up in Singapore Towards healthy Outcomes (GUSTO) birth cohort study. Int J Epidemiol. 2014 Oct;43(5):1401-9. DOI: 10.1093/ije/dyt125.
- 3. Agency for Science, Technology and Research. GUSTO Data vault [Internet]. Singapore: Agency for Science, Technology and Research; [cited 2023 Apr 10]. Available from: <u>https://gustodatavault.sg/</u>.
- 4. Ho C, Ang LT, Ng M, Png H, Tan SL, Ye E, et al. GUSTO Data Vault: Working Towards OMOP Data Standardisation. Poster session presented at: 2022 APAC OHDSI Symposium; 2022 Nov 13; Taiwan.
- Observational Health Data Sciences and Informatics. OMOP Common Data Model [Internet]. [place unknown: Observational Health Data Sciences and Informatics]; [cited 2023 Apr 10]. Available from: <u>https://ohdsi.github.io/CommonDataModel/cdm54.html.</u>
- Wilkinson M, Dumontier M, Aalbersberg I, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3 [Internet]. 2016 Mar [cited 2023 Apr 10]; 160018 (2016). Available from: <u>https://www.nature.com/articles/sdata201618</u> DOI: 10.1038/sdata.2016.18